Direct search using probabilistic descent

Clément Royer
ENSEEIHT-IRIT, University of Toulouse, France

Co-authors: S. Gratton, L. N. Vicente, Z. Zhang

ISMP 2015, Pittsburgh
July 16, 2015

Outline

Introductory assumptions and definitions

We consider an unconstrained smooth problem:

$$\min_{x \in \mathbb{R}^n} f(x).$$

Assumptions on f

- f bounded from below;
- ∇f Lipschitz continuous of constant ν.

Introductory assumptions and definitions

We consider an unconstrained smooth problem:

$$\min_{x \in \mathbb{R}^n} f(x).$$

Assumptions on f

- f bounded from below;
- ∇f Lipschitz continuous of constant ν.

Solving the problem using the derivative

- **Steepest descent method:** move in the direction of $-\nabla f(x)$;
- **Gradient-related methods:** use descent directions that make an acute angle with $-\nabla f(x)$.

Derivative-Free Optimization

Derivative-Free Optimization (DFO) methods

- Assume that the gradient is unavailable (Ex: simulation code);
- Two main (deterministic) classes:
  - Model-based methods;
  - Direct-search methods.

Introduction to Derivative-Free Optimization
A.R. Conn, K. Scheinberg, L.N. Vicente. (2009)

Derivative-Free Optimization

Derivative-Free Optimization (DFO) methods

- Assume that the gradient is unavailable (Ex: simulation code);
- Two main (deterministic) classes:
  - Model-based methods;
  - Direct-search methods.

Introduction to Derivative-Free Optimization
A.R. Conn, K. Scheinberg, L.N. Vicente. (2009)

In this talk, we look at directional direct-search methods.

Optimization by Direct Search: New Perspectives on Some Classical and Modern Methods
T.G. Kolda, R.M. Lewis, V. Torczon (2003).

Outline

A basic framework for direct-search algorithms

1. **Initialization:** Set $x_0, \alpha_0, \theta < 1 \leq \gamma$.
2. **For** $k = 0, 1, 2, \ldots$
   - Choose a set D_k of m vectors.
   - If it exists $d_k \in D_k$ so that

     $$f(x_k + \alpha_k d_k) < f(x_k) - \alpha_k^2,$$

     then declare k *successful*, set $x_{k+1} := x_k + \alpha_k d_k$ and update $\alpha_{k+1} := \gamma \alpha_k$.
   - Otherwise declare k *unsuccessful*, set $x_{k+1} := x_k$ and update $\alpha_{k+1} := \theta \alpha_k$.

A basic framework for direct-search algorithms

1. **Initialization:** Set $x_0, \alpha_0, \theta < 1 \leq \gamma$.
2. **For** $k = 0, 1, 2, \ldots$
   - Choose a set D_k of m vectors.
   - If it exists $d_k \in D_k$ so that

     $$f(x_k + \alpha_k d_k) < f(x_k) - \alpha_k^2,$$

     then declare k *successful*, set $x_{k+1} := x_k + \alpha_k d_k$ and update $\alpha_{k+1} := \gamma \alpha_k$.
   - Otherwise declare k *unsuccessful*, set $x_{k+1} := x_k$ and update $\alpha_{k+1} := \theta \alpha_k$.

Polling choice in deterministic direct search

We would like to choose directions/polling sets D_k sufficiently good to ensure convergence:

- How do we know that a set is good ?
- What is the role of this quality in the convergence proof ?

Polling choice in deterministic direct search

We would like to choose directions/polling sets D_k sufficiently good to ensure convergence:

- How do we know that a set is good ?
- What is the role of this quality in the convergence proof ?

A measure of set quality

For a set of vectors D,

$$\mathrm{cm}(D) = \min_{v \in \mathbb{R}^n \setminus \{0\}} \max_{d \in D} \frac{d^\top v}{\|d\| \, \|v\|}$$

is the cosine measure of D.

Polling choice in deterministic direct search

We would like to choose directions/polling sets D_k sufficiently good to ensure convergence:

- How do we know that a set is good ?
- What is the role of this quality in the convergence proof ?

A measure of set quality

For a set of vectors D,

$$\mathrm{cm}(D) = \min_{v \in \mathbb{R}^n \setminus \{0\}} \max_{d \in D} \frac{d^\top v}{\|d\| \, \|v\|}$$

is the cosine measure of D.

A good cosine measure

When $\mathrm{cm}(D) > 0$, any vector makes an acute angle with an element of D.

Set quality

We would like to have $\mathrm{cm}(D) > 0$.
A common choice is to use positive spanning sets.

Positive Spanning Sets (PSS)

D is a PSS if it generates $\mathbb{R}^n$ by nonnegative linear combinations.

- D is a PSS **iff** $\mathrm{cm}(D) > 0$;
- a PSS contains at least $n + 1$ vectors.

Set quality

We would like to have $\mathrm{cm}(D) > 0$.
A common choice is to use positive spanning sets.

Positive Spanning Sets (PSS)

D is a PSS if it generates $\mathbb{R}^n$ by nonnegative linear combinations.

- D is a PSS **iff** $\mathrm{cm}(D) > 0$;
- a PSS contains at least $n + 1$ vectors.

Example

$D_{\oplus} = [I \ -I]$ is a PSS with

$$\mathrm{cm}\left(D_{\oplus}\right) = \frac{1}{\sqrt{n}}.$$

Convergence for deterministic direct search

Lemma

$$\lim_{k \to \infty} \alpha_k = 0.$$

Proposition

If the k-th iteration is unsuccessful and $\mathrm{cm}(D_k) \geq \kappa > 0$*, then*

$$\kappa \, \|\nabla f(x_k)\| \leq \mathcal{C}(\nu) \, \alpha_k.$$

Convergence for deterministic direct search

Lemma

$$\lim_{k \to \infty} \alpha_k = 0.$$

Proposition

If the k-th iteration is unsuccessful and $\mathrm{cm}(D_k) \geq \kappa > 0$, *then*

$$\kappa \, \|\nabla f(x_k)\| \leq \mathcal{C}(\nu) \, \alpha_k.$$

Convergence Theorem

If $\forall k, \ \mathrm{cm}(D_k) \geq \kappa$, we have

$$\liminf_{k \to \infty} \|\nabla f(x_k)\| = 0.$$

Worst-case complexity in deterministic direct search

Theorem (Vicente '13)

Let N_ϵ be the number of function evaluations needed to reach an point such that $\|\nabla f(x_k)\| < \epsilon$; then

$$N_\epsilon \leq \mathcal{O}\left(m\left(\kappa\,\epsilon\right)^{-2}\right).$$

Corollary (Vicente '13, Konečny and Richtárik '14)

Choosing $D_k = D_\oplus$, one has $\kappa = 1/\sqrt{n}, m = 2n$, and the bound becomes

$$N_\epsilon \leq \mathcal{O}\left(n^2\,\epsilon^{-2}\right).$$

The bound is optimal in terms of powers of n.
(Dodangeh, Vicente, Zhang '14)

Outline

Introducing randomness

Idea (Gratton and Vicente '13)

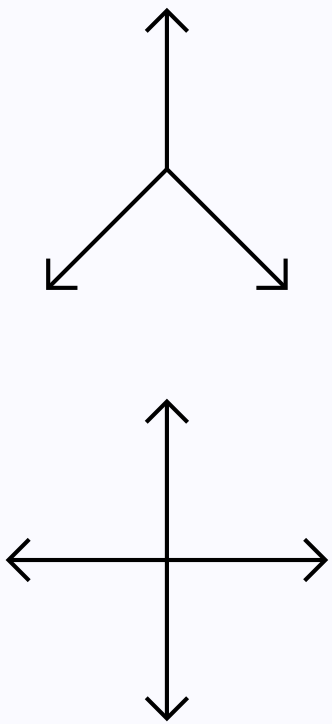
Randomly independently generate polling sets, possibly of less than $n+1$ vectors!

Introducing randomness

Idea (Gratton and Vicente '13)

Randomly independently generate polling sets, possibly of less than $n + 1$ vectors!

From PSS...

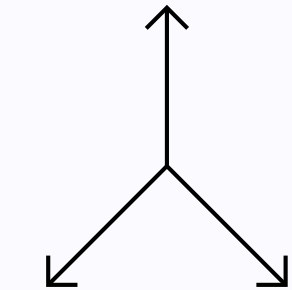


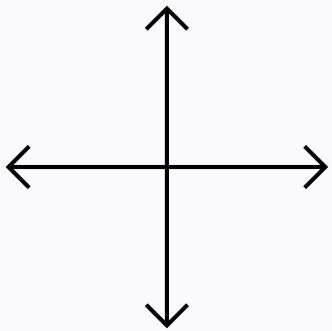
Introducing randomness

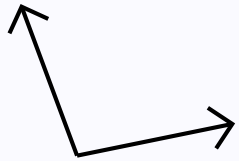
Idea (Gratton and Vicente '13)

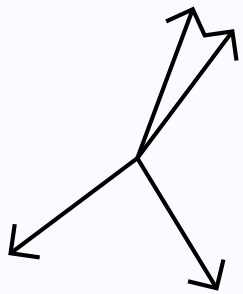
Randomly independently generate polling sets, possibly of less than $n+1$ vectors!

From PSS...









...to random sets

Numerical motivations

- Some results for $n = 40$ and $\epsilon = 10^{-3}$:

Problem	$[I\ -I]$	$[Q_k\ -Q_k]$	$2\,n$	$n+1$	$n/2$	2	1
arglina	3.42	16.67	10.30	6.01	3.21	1.00	–
arglinb	20.50	11.38	7.38	2.81	2.35	1.00	2.04
broydn3d	4.33	11.22	6.54	3.59	2.04	1.00	–
dqrtic	7.16	19.50	9.10	4.56	2.77	1.00	–
engval1	10.53	23.96	11.90	6.48	3.55	1.00	2.08
freuroth	56.00	1.33	1.00	1.67	1.33	1.00	4.00
integreq	16.04	18.85	12.44	6.76	3.52	1.00	–
nondquar	6.90	17.36	7.56	4.23	2.76	1.00	–
sinquad	–	2.12	1.31	1.00	1.60	1.23	–
vardim	1.00	3.30	1.80	2.40	2.30	1.80	4.30

Table : Relative number of function evaluations for different types of polling (mean on 10 runs)

A probabilistic direct-search algorithm

From deterministic to probabilistic notations

- Polling sets : $D_k = \mathfrak{D}_k(\omega)$;
- Iterates : $x_k = X_k(\omega)$;
- Step sizes : $\alpha_k = \mathcal{A}_k(\omega)$.

1. **Initialization:** Set $x_0, \alpha_0, \theta < 1 \leq \gamma$.
2. **For** $k = 0, 1, 2, ...,$
   - Choose a set $\mathfrak{D}_k$ of m independent random vectors.
   - If it exists $\mathfrak{d}_k \in \mathfrak{D}_k$ so that

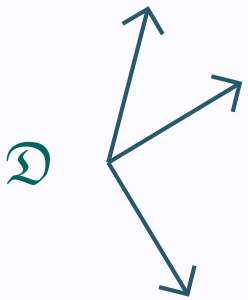
$$f(X_k + \mathcal{A}_k \mathfrak{d}_k) < f(X_k) - \mathcal{A}_k^2,$$

     then declare k successful, set $X_{k+1} := X_k + \mathcal{A}_k \mathfrak{d}_k$ and update $\mathcal{A}_{k+1} := \gamma \mathcal{A}_k$.
   - Otherwise, declare k unsuccessful, set $X_{k+1} := X_k$ and update $\mathcal{A}_{k+1} := \theta \mathcal{A}_k$.

Outline

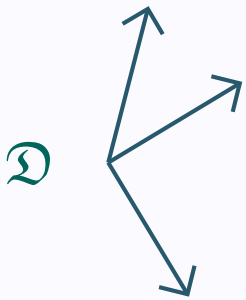
A new measure of set quality

- $\mathfrak{D}$ is not a PSS...

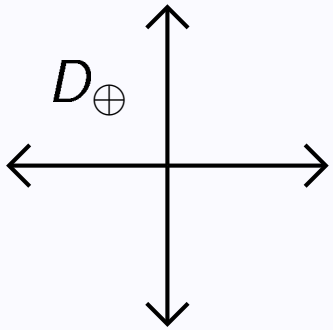


A new measure of set quality

$\mathfrak{D}$ is not a PSS...

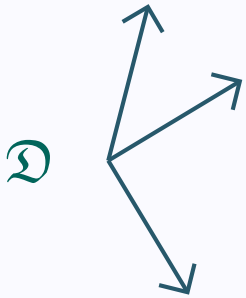


...$D_\oplus$ is...

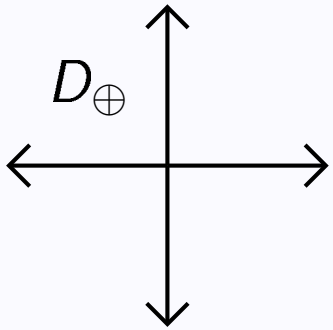


A new measure of set quality

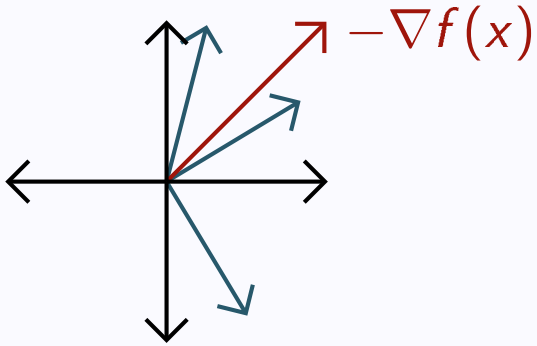
$\mathfrak{D}$ is not a PSS...



...$D_\oplus$ is...



...but here $-\nabla f(x)$ is closer to $\mathfrak{D}$!



A new measure of set quality

Set assumption in the deterministic case

- We required:

$$\mathrm{cm}(D_k) = \min_{v \neq 0} \max_{d \in D_k} \frac{d^\top v}{\|d\| \, \|v\|} \geq \kappa.$$

- Yet we only used:

$$\mathrm{cm}\left(D_k, -\nabla f(x_k)\right) = \max_{d \in D_k} \frac{d^\top (-\nabla f(x_k))}{\|d\| \|\nabla f(x_k)\|} \geq \kappa.$$

- In the random case, the second one might happen with some probability;
- The main issue is to find the adequate probabilistic tools to express this fact.

(p, κ)-descent sets

- We want to look at $\mathbb{P}\left(\mathrm{cm}\left(\mathfrak{D}_k, -\nabla f(X_k)\right) \geq \kappa\right)$, but X_k depends on $\mathfrak{D}_0, \ldots, \mathfrak{D}_{k-1}$;
- A solution is conditioning to the past.

Convergence on trust-region methods based on probabilistic models
A.S. Bandeira, K. Scheinberg, L.N. Vicente. (2014)

(p, κ)-descent sets

- We want to look at $\mathbb{P}\left(\mathrm{cm}\left(\mathfrak{D}_k, -\nabla f(X_k)\right) \geq \kappa\right)$, but X_k depends on $\mathfrak{D}_0, \ldots, \mathfrak{D}_{k-1}$;
- A solution is conditioning to the past.

Convergence on trust-region methods based on probabilistic models
A.S. Bandeira, K. Scheinberg, L.N. Vicente. (2014)

Probabilistic descent property

A random set sequence $\{\mathfrak{D}_k\}$ is said to be (p, κ)-descent if:

$$\begin{aligned} \mathbb{P}\left(\mathrm{cm}\left(\mathfrak{D}_0, -\nabla f(x_0)\right) \geq \kappa\right) &\geq p \\ \forall k \geq 1, \quad \mathbb{P}\left(\mathrm{cm}\left(\mathfrak{D}_k, -\nabla f(X_k)\right) \geq \kappa \,\middle|\, \mathfrak{S}_{k-1}^{\mathfrak{D}}\right) &\geq p, \end{aligned}$$

where $\mathfrak{S}_{k-1}^{\mathfrak{D}} = \sigma(\mathfrak{D}_0, \ldots, \mathfrak{D}_{k-1})$.

Convergence results

Lemma

For all realizations $\{\alpha_k\}$ *of* $\{\mathcal{A}_k\}$*:*

$$\lim_{k\to\infty} \alpha_k = 0.$$

Almost-sure Convergence Theorem

If $\{\mathfrak{D}_k\}$ is (p, κ)-descent with $p > \ln(\theta)\,\ln(\theta/\gamma)^{-1}$, then

$$\mathbb{P}\left(\liminf_{k\to\infty} \|\nabla f(X_k)\| = 0 \right) = 1.$$

Sketch of the proof

Two main arguments:

Lemma

If k is an unsuccessful iteration; then

$$\{\operatorname{cm}(\mathfrak{D}_k, -\nabla f(X_k)) \geq \kappa\} \subset \{\kappa \|\nabla f(X_k)\| \leq \mathcal{C}(\nu)\mathcal{A}_k\}.$$

Proposition

Let $Z_k = \mathbf{1}(\operatorname{cm}(\mathfrak{D}_k, -\nabla f(X_k)) \geq \kappa)$; then

$$S_k = \sum_{i=0}^{k-1} \left[\left(1 - \frac{\ln \gamma}{\ln \theta}\right) \cdot Z_i - 1\right].$$

is a submartingale and $\mathbb{P}(\limsup S_k = \infty) = 1$.

A practical (p, κ)-descent sequence

In order to ensure global convergence, one must assume:

$$p > p_0 = \frac{\ln(\theta)}{\ln(\theta/\gamma)}.$$

This can give a lower bound on $m = |\mathfrak{D}_k|$.

A practical example: uniform distribution over the unit sphere

If

$$m > \log_2 \left(1 - \frac{\ln \theta}{\ln \gamma}\right),$$

then there exist p and τ totally determined by γ and θ such that the sequence $\mathfrak{D}_k$ is $(p, \tau/\sqrt{n})$-descent, with $p > p_0$.

If $\gamma = \theta^{-1} = 2$, it suffices to choose $m \geq 2$.

Outline

WCC for probabilistic descent

Intuitive idea

Let $G_k = \nabla f(X_k)$, so $Z_k = \mathbf{1}\left(\mathrm{cm}(\mathfrak{D}_k, -G_k) \geq \kappa\right)$.

- If $Z_k = 1$ and k unsuccessful, then $\|G_k\| < \mathcal{O}(\mathcal{A}_k)$...

WCC for probabilistic descent

Intuitive idea

Let $G_k = \nabla f(X_k)$, so $Z_k = \mathbf{1}\left(\operatorname{cm}(\mathfrak{D}_k, -G_k) \geq \kappa\right)$.

- If $Z_k = 1$ and k unsuccessful, then $\|G_k\| < \mathcal{O}(\mathcal{A}_k)$...
- ...so if $\inf_{0 \leq l \leq k} \|G_l\|$ has not decreased much, $\sum_{l=0}^{k} Z_l$ should not be too high.

WCC for probabilistic descent

Intuitive idea

Let $G_k = \nabla f(X_k)$, so $Z_k = \mathbf{1}\left(\mathrm{cm}(\mathfrak{D}_k, -G_k) \geq \kappa\right)$.

- If $Z_k = 1$ and k unsuccessful, then $\|G_k\| < \mathcal{O}(\mathcal{A}_k)$...
- ...so if $\inf_{0 \leq l \leq k} \|G_l\|$ has not decreased much, $\sum_{l=0}^{k} Z_l$ should not be too high.

A useful bound

For all realizations of the algorithm, one has

$$\sum_{l=0}^{k} z_l \;\leq\; \mathcal{O}\left(\frac{1}{\kappa^2 \, \|\tilde{g}_k\|^2}\right) + p_0 \, k,$$

with $\|\tilde{g}_k\| = \inf_{0 \leq l \leq k} \|g_l\|$.

WCC for probabilistic descent (2)

An inclusion argument

$$\left\{ \|\tilde{G}_k\| \geq \epsilon \right\} \subset \left\{ \sum_{l=0}^{k} Z_l \leq \lambda k \right\}$$

with $\lambda = \mathcal{O}\left(\frac{1}{k\,\kappa^2\,\epsilon^{-2}}\right) + p_0$.

A Chernoff-type probability result

For any $\lambda \in (0, p)$,

$$\mathbb{P}\left(\sum_{l=0}^{k-1} Z_l \leq \lambda k \right) \leq \exp\left[-\frac{(p-\lambda)^2}{2\,p} k \right].$$

WCC for probabilistic descent (3)

Probabilistic worst-case complexity

Let $\{\mathfrak{D}_k\}$ be (p, κ)-descent, $\epsilon \in (0, 1)$ and N_ϵ the number of function evaluations needed to have $\|\tilde{G}_k\| \leq \epsilon$. Then

$$\mathbb{P}\left(N_\epsilon \leq \mathcal{O}\left(\frac{m\,(\kappa\epsilon)^{-2}}{p - p_0}\right)\right) \geq 1 - \exp\left(-\mathcal{O}\left(\frac{p - p_0}{p}\epsilon^{-2}\right)\right).$$

WCC for probabilistic descent (3)

Probabilistic worst-case complexity

Let $\{\mathfrak{D}_k\}$ be (p, κ)-descent, $\epsilon \in (0, 1)$ and N_ϵ the number of function evaluations needed to have $\|\tilde{G}_k\| \leq \epsilon$. Then

$$\mathbb{P}\left(N_\epsilon \leq \mathcal{O}\left(\frac{m\,(\kappa\epsilon)^{-2}}{p - p_0}\right)\right) \geq 1 - \exp\left(-\mathcal{O}\left(\frac{p - p_0}{p}\epsilon^{-2}\right)\right).$$

- By taking $\mathfrak{D}_k = D_\oplus$, one has $\kappa = 1/\sqrt{n}$, $m = 2\,n$ and $p = 1$, we recover:

$$\mathcal{O}\left(n^2\,\epsilon^{-2}\right).$$

- With uniform generation, one can decrease this rate to $\mathcal{O}(m\,n\,\epsilon^{-2})$, with $m \ll n + 1$!

Summary

What comes out from our study ?

- An almost surely convergent method that does not rely on PSS;

Summary

What comes out from our study ?

- An almost surely convergent method that does not rely on PSS;
- A new probabilistic worst-case complexity argument, adaptable to other DFO methods (e.g. Trust-Region);

Summary

What comes out from our study ?

- An almost surely convergent method that does not rely on PSS;
- A new probabilistic worst-case complexity argument, adaptable to other DFO methods (e.g. Trust-Region);
- Improved numerical performance.

To be continued

The paper

Direct Search based on Probabilistic Descent.
S. Gratton, C. W. Royer, L. N. Vicente, Z. Zhang.
To appear in SIAM Journal on Optimization.

What is next ?

- Second-order results and probabilistic assumptions;
- Extension to nonsmooth/constrained problems.

Thank you for your attention !